



TIDES

*Translingual Information
Detection, Extraction, and
Summarization*



Why TIDES?

- 200M Web pages as of 7/97
- 1 M terabytes of audio / year
- Uncounted printed matter
- Foreign language information growing faster than English



National Security

தலைமைச் செயல்கம்

தமிழ்மீடு விடுதலைப் புலிகள்

தமிழ்மூம்

13.05.1998

எமது தேசிய விடுதலைப் போராட்ட வரலாற்றில் இன்முக்கியத்துவம் வாய்ந்த நாள். எமது எதிரியான சிறிமிகப் பெரிய படையெடுப்பான “ஸ்ரீயசிக்குரு” இராஷ்டிரித்து நின்று போராடி, இன்றுள்ள ஒராண்டு பூர்த்தமாத காலத்திற்குள் முடிந்துவிடுமென போர்ப்பறை அபிரச்சார எடுப்புடன் ஆரம்பமான இப்போர் நடவடிக்கவுடமாகியும் இன்னும் முடிவுபெறாது இழுபடுகிறது. எட்டிவிட்ட ஒரு தனிச்சமர் என்ற ரீதியில், தமிழ்மூப் வரலாற்றில் மட்டுமின்றி உலகப் போரியல் வரலாற்று நின்டதொரு சமராக இது முக்கியத்துவம் பெறுகிறது. படையெடுப்பை மூர்க்கமாக எதிர்த்துப் போராடி, எதிராண்மூலானாக நகர்வு வேகத்தை தடுத்து நிறுத்தி, எதிரிப்படைகளை வன்னிக்காட்டிற்குள் முடக்கி வைத்து உலக இராணுவ வராலற்றில் ஒரு ஒப்பற்ற சாதனங்கை எமது விடுதலை இயக்கம் நிலை நாட்டியிக்கிறது.



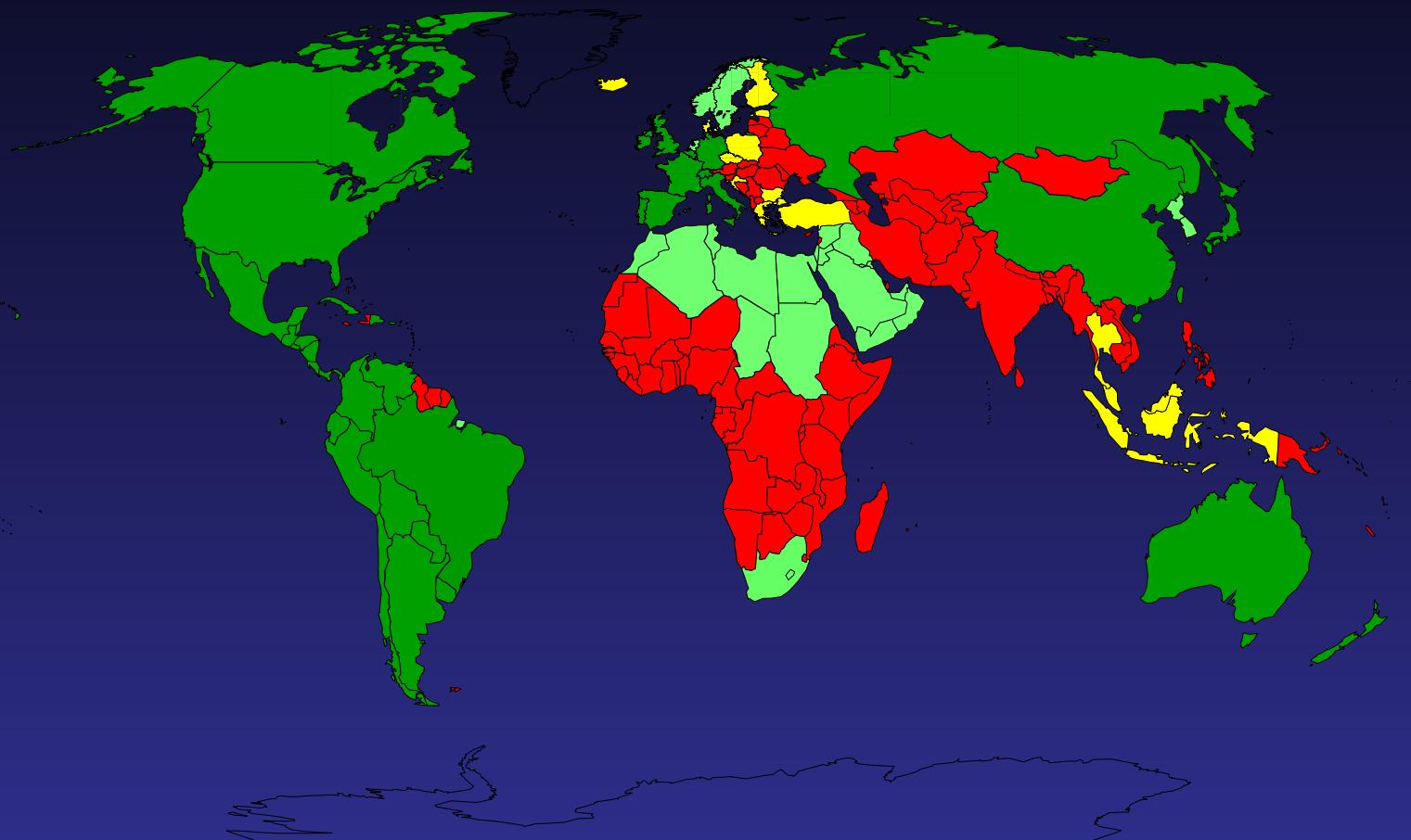


TIDES Goal

- Find and Interpret Information Vital to National Security
 - Retrieve unfamiliar languages
 - Translate into English
 - Extract and correlate content



Machine Translation



ITC



Bombs & Warnings





Targets

- Translingual access rivaling monolingual access
- Rapid development of MT for new languages
- Multi-document information extraction and correlation



The Problem

- Exhaustive coverage expected
- Many simmering pots
- Unpredictable flare-ups
- Accurate analysis critical

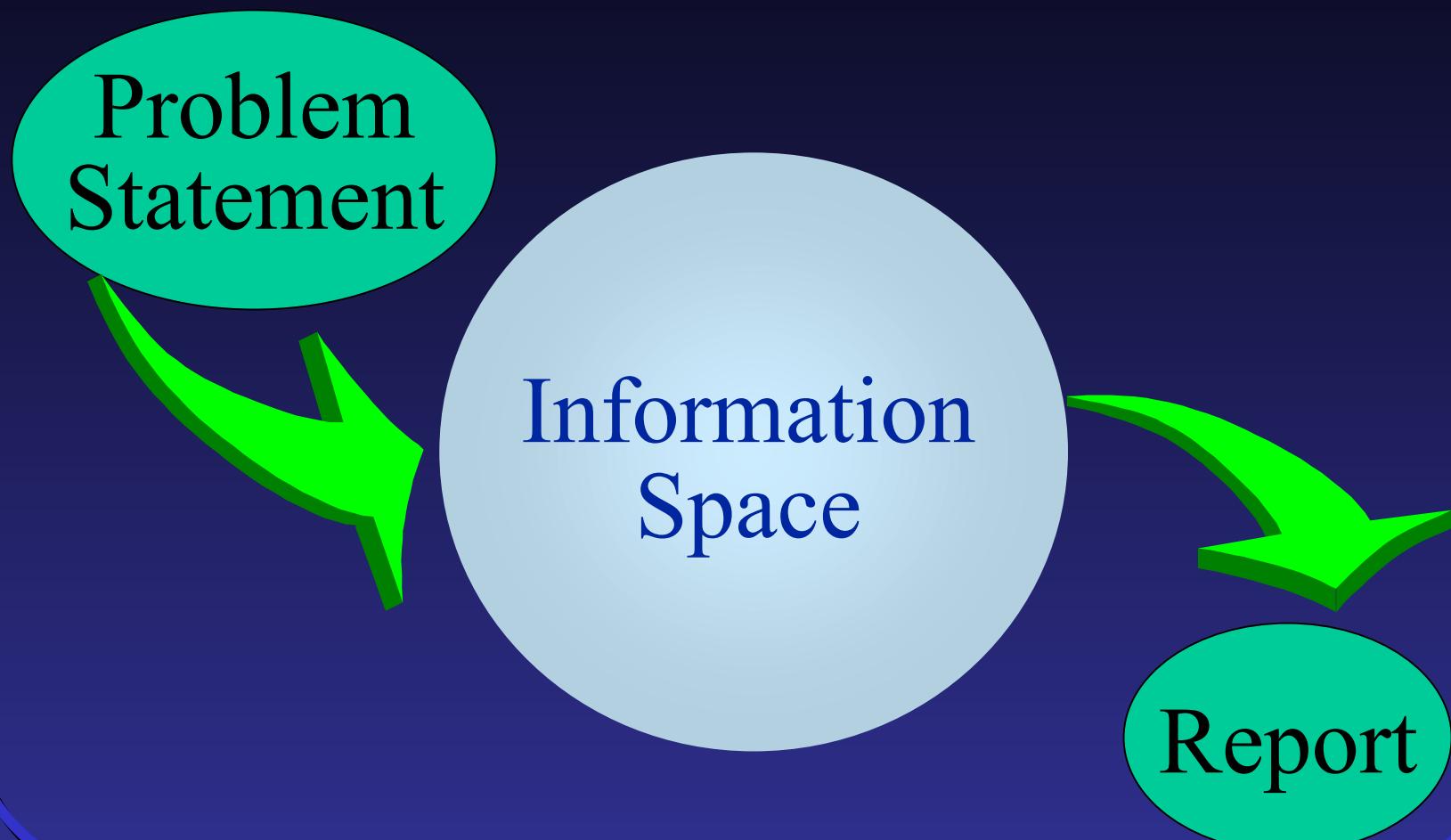


The World - 1999

- ~228 Countries
- >6,700 Languages
- >39,000 Language, dialect,
and alternate names



Framework





Process Steps

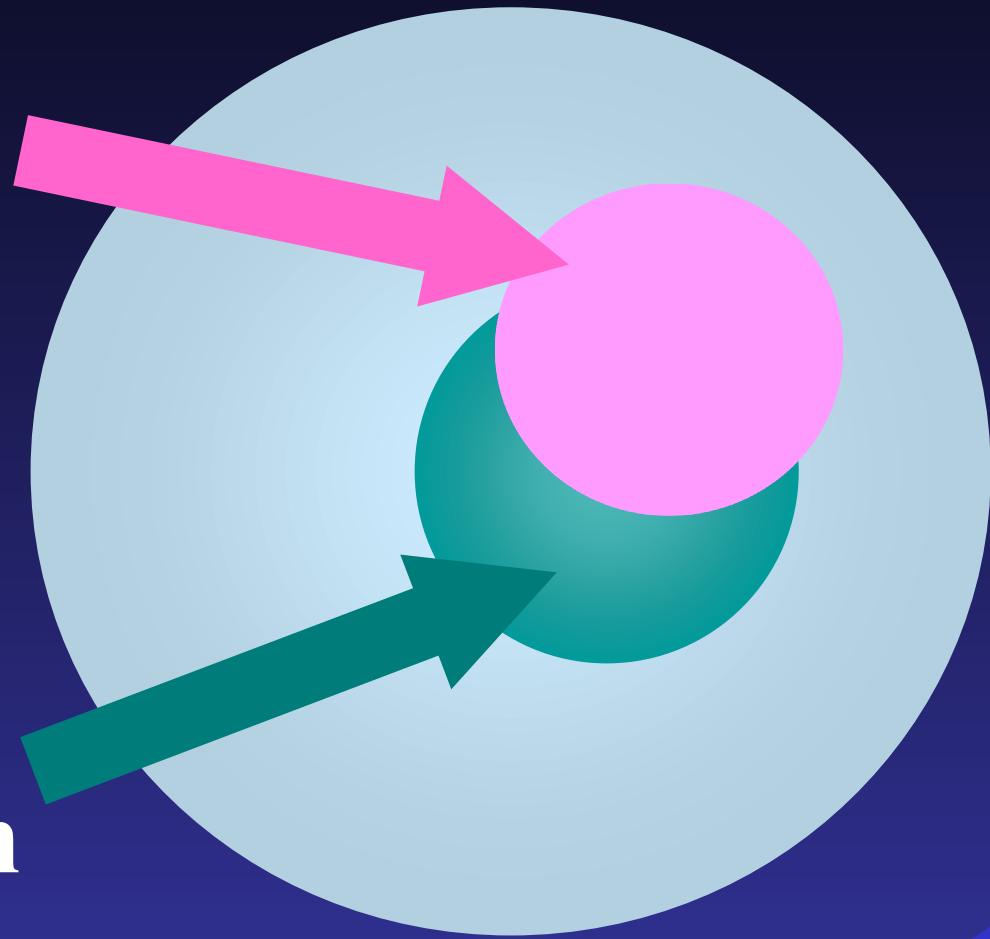
- Information Retrieval
 - Topic Detection
 - Entity Extraction
 - Summary Generation



Information Retrieval

Retrieved
Information

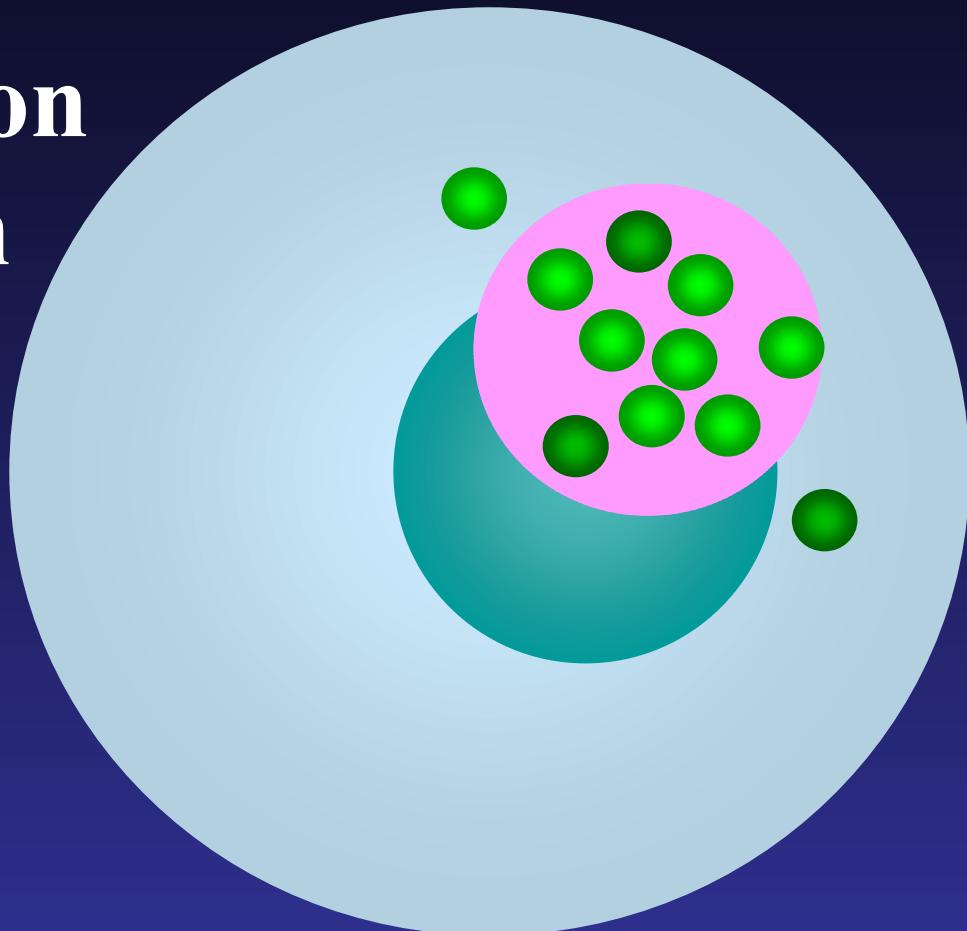
Relevant
Information





Topic Detection

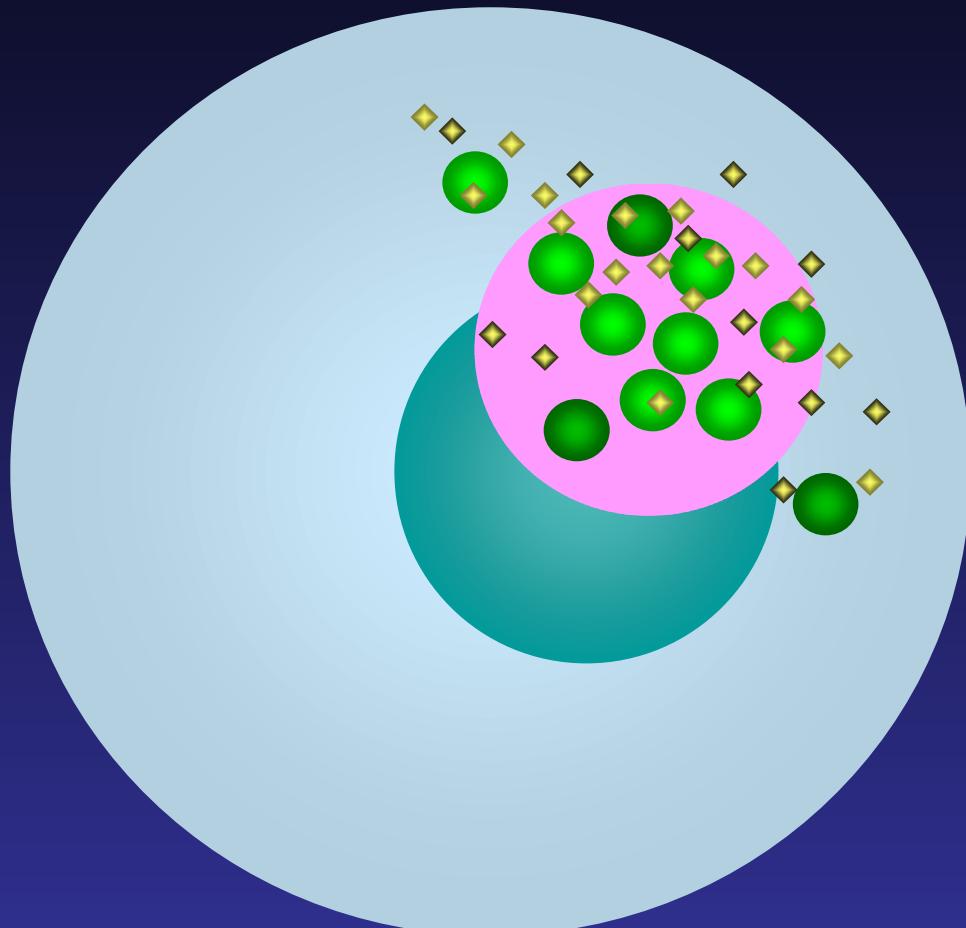
- Segmentation
- Recognition
- Tracking





Entity Extraction

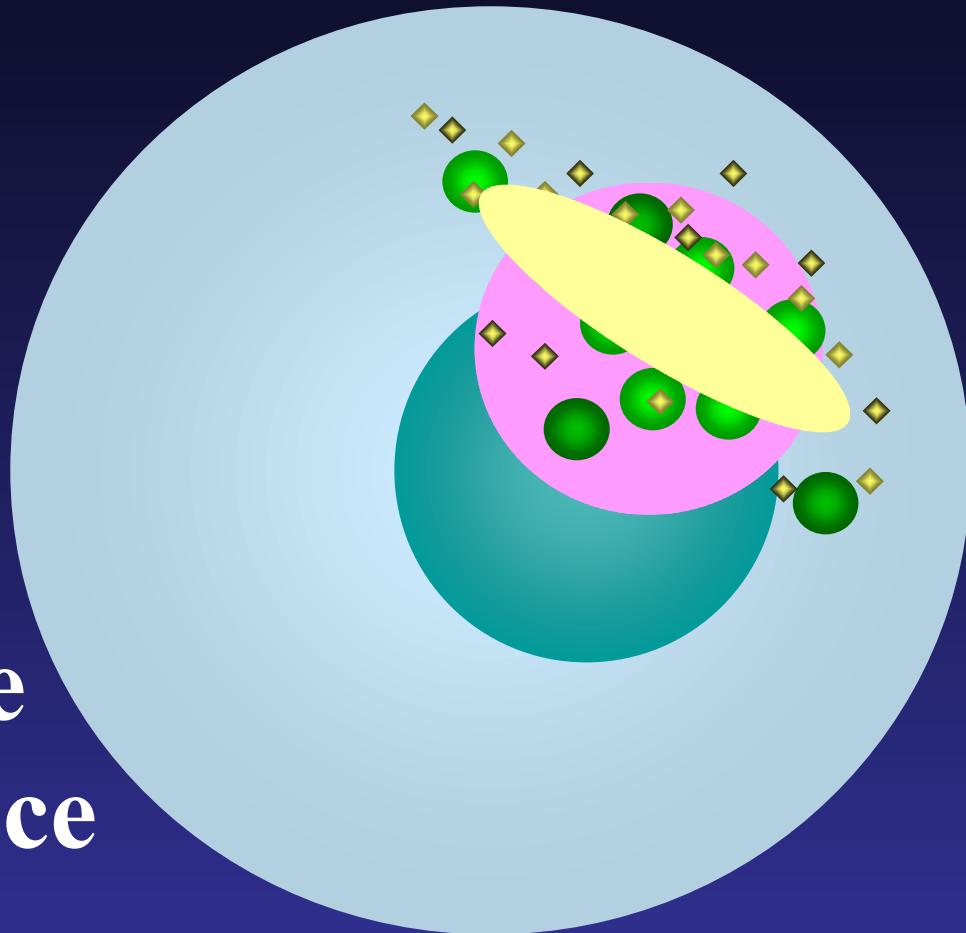
- Names
- Places
- Events





Summarization

- Type
- Content
- Perspective
- Performance





Environment

- Large information space
- Human knowledge, patience, and labor
- Relevance feedback
- Monolingual (English)



Beyond English

- Query translation
- Document translation
- 50% performance of monolingual retrieval



Exploiting Feedback

- Relevance feedback
- Topic unification
- Content threading
- Multidocument summarization



3-Year Goals

- Improved translingual IR
- Rapid shift to new language
- Multilingual topic recognition
- Multidocument summarization



5-Year Goals

- 30+ languages
- Multilingual entity correlation
- Multilingual templates
- Multilingual summarization



TIDES

The collage on the left includes:

- CBS RADIO NETWORK logo
- TV screens showing CNN, ABC, CBS News, NBC, and BBC
- A hand holding a mobile phone
- A "CHAT ROOM" icon
- A newspaper clipping from THE WALL STREET JOURNAL
- A globe surrounded by a network of icons
- A portable radio device
- A newspaper clipping from The New York Times
- An envelope icon at the bottom

The central illustration shows a suspension bridge spanning a body of water, with large brown landmasses on either side.

The five software windows on the right are:

- SUMMARY**: Shows a document with dense text.
- PHOTOS**: Shows a camera icon.
- CHARTS**: Shows a line graph with a blue line fluctuating.
- BACKGROUND**: Shows a grayscale image of a landscape.

• Translingual Access

• Machine Translation

• Summarization and Correlation

ITC